Instructions for Reviewing Your Edited Manuscript

Reviewing the edits

- **Do not generate a new electronic version of your manuscript.**
- Indicate your changes in the file using the Adobe Comment tool. **Do not make any changes directly in the file using strikethroughs and underscores**. Upload the annotated PDF file to the JAMA Network production system using the encrypted link sent to you in the email.
- If you are unable to use Adobe, indicate all changes by page number and line number. Please include the entire sentence, with your changes clearly marked (in bold, italics, or underline).

Tables and figures

- If your manuscript contains tables, review them for content only, not for bad word breaks, the shapes of the cells, or formatting.
- If you are asked to provide additional data throughout a table, you may upload a separate document that contains the complete information rather than mark each cell using Adobe Comment tools.
- If your manuscript has figures, they are being formatted and will be sent to you separately or included in the proof that we will send to you for final review.

Confidentiality, media attention, and disclosures, and reprints

- The information in your manuscript is confidential. You should not distribute copies of the edited manuscript to anyone, except coauthors, without approval from the editorial office. The news media should not release any information about your article until the date of publication. If you need help coordinating news releases or media attention, contact JAMA Network Media Relations at mediarelations@jamanetwork.org or 312-464-5262.
- Our policy requires that all authors disclose all potential conflicts of interest, including specific financial interests and relationships and affiliations (other than those listed in the author affiliations). Definitions and terms of such disclosures can be found at the journal's Instructions for Authors. Authors should err on the side of full disclosure and should contact the editorial office if they have questions or concerns. Please verify and confirm that all Conflict of Interest disclosure information for you and all coauthors is accurate, complete, up-to-date, and reported accurately and in its entirety.
- You may order reprints or eprints online at www.ama-authorreprints.com.

Sincerely yours,
Emma Hyche
Manuscript Editor
JAMA Network
emma.hyche@jamanetwork.org

Original Investigation

Deep Learning Phenotyping of Tricuspid Regurgitation for Automated ~~High Throughput~~ Transthoracic Echocardiography~~TTE~~ Assessment

Amey Vrudhula, MD1,2

Milos Vukadinovic, BS1,3

Christiane Haeffele, MD4

Alan C. Kwan, MD1

Daniel Berman, MD1

David Liang, MD, PhD4

Robert Siegel, MD

Susan Cheng, MD, MMSc, MPH1

David Ouyang, MD1,5

[1]Smidt Heart Institute, Department of Cardiology~~, Smidt Heart Institute~~, Cedars-Sinai Medical Center, Los Angeles, California

[2]Icahn School of Medicine at Mt Sinai, New York, New York

[3]Department of Bioengineering, University of California, Los Angeles~~, Los Angeles, California~~

Commented [EH1]: I've trimmed this title down slightly to get it closer to JAMA's character limit for this article type (100, including spaces). Let me know if you have any alternate or additional suggestions to get the title closer to this limit!

Commented [EH2]: Please review the author names and verify that the first (given) names (including initials) and surnames are correctly tagged. First names (including initials) are highlighted in blue, surnames are highlighted in green.

Commented [EH3]: Please provide an affiliation for Dr Siegel

[4]Division of Cardiology, Department of Medicine~~, Division of Cardiology~~, Stanford University, Palo Alto, California

[5]Division of Artificial Intelligence in Medicine, Cedars-Sinai Medical Center, Los Angeles, California

Question: Can an automated deep learning workflow accurately assess presence and severity of tricuspid regurgitation (TR) from full transthoracic echocardiography (TTE) studies at a large scale?

Findings: In geographically and temporally distinct test cohorts, the deep learning workflow showed strong and consistent performance in identifying apical 4-chamber~~A4C~~ videos with color Doppler across the tricuspid valve (area under the receiver operating characteristic curve [AUC], ≥0.999 [~~(~~0.998-~~—~~0.999]~~)~~) and in detecting studies with clinically significant TR (AUC, ≥0.928 [~~(~~0.913-~~—~~0.943]~~)~~).

Meaning: In this study, an end-to-end deep learning pipeline automatically identified clinically significant TR from thousands of full TTE studies with excellent performance.

Importance: Accurate assessment of tricuspid regurgitation (TR) is necessary for identification and risk stratification.

Objective: To design a deep learning computer vision workflow for identifying color Doppler echocardiogram videos and characterizing~~e~~ TR severity.

Design, Setting, and Participants: An automated deep learning workflow was developed using 47 312 studies (2 079 898 videos) from Cedars-Sinai Medical Center (CSMC) between 2011

Commented [EH4]: Are these median (IQR) values? 95% confidence intervals? We'll need some more information here. Depending on what these data are, we may need to omit them from the Key Points, which should not contain hard data.

and 2021. The pipeline was tested on a temporally distinct test set of 2462 studies (108 138 videos) obtained in 2022 at CSMC and a geographically distinct cohort of 5549 studies (278 377 videos) from Stanford Healthcare (SHC).

~~**Setting:** Cohorts from CSMC and SHC.~~

~~**Participants:**~~ Training and validation cohorts contained data from 31 708 patients at CSMC ~~patients~~ receiving care between 2011 and 2021. Patients were chosen for parity across TR severity classes, with no exclusion criteria based on other clinical or demographic characteristics. The 2022 CSMC test cohort and SHC test cohort contained studies of 2132 patients and 4798 patients, respectively.

Exposure:

Main Outcome~~(s)~~ and Measure~~(s)~~: The main outcomes were area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity in identifying apical 4-chamber (A4C) videos with color Doppler across the tricuspid valve and AUROC in identifying studies with moderate-to-severe or severe TR.

Results: Of XX total patients, XX patients (XX%) were female. In the CSMC test dataset, the view classifier demonstrated an AUC of 1.000 (0.999-~~—~~1.000) and identified at least 1~~one~~ A4C video with color Doppler across the tricuspid valve with a sensitivity of 0.975 (0.968-0.982) and a specificity of 1.000 (1.000-1.000). In the CSMC test cohort, moderate-to~~or~~-severe TR was detected with an AUC of 0.928 (0.913-~~—~~0.943), and severe TR was detected with an AUC of 0.956 (0.940-~~—~~0.969). In the SHC cohort, the view classifier correctly identified at least 1~~one~~

Commented [EH5]: We'll need to include when the period of data analysis was (ie, when the workflow was designed and tested).

Commented [EH6]: Please provide a short phrase or sentence summarizing the study's exposure, if applicable.

Commented [EH7]: This is referred to as AUC in the Figures and elsewhere in the manuscript; which version would you prefer?

Commented [EH8]: This section usually contains some demographic info, like No. (%) female and mean (SD) age. Since the latter of these doesn't seem relevant, we'll keep just the former. Could you provide this missing info?

Commented [EH9]: See prior query: AUC or AUROC?

Commented [EH10]: See Key Points query. I'm assuming the values in parentheses are 95% CIs, but I'd like to confirm that.

Commented [EH11]: Is this missing the number of studies in which this was true and the total number, like in the SHC cohort sentence below?

Commented [EH12]: I'm assuming this should be "moderate-to-severe," like in the Main Outcomes and Measures section; if it should be "moderate or severe," just let me know

TR color Doppler video in 5268 of the 5549 studies, resulting in an AUC of 0.999 (0.998-—0.999), a sensitivity of 0.949 (0.944-—0.955), and a specificity of 0.999 (0.999-—0.999). The artificial intelligence~~AI~~ model detected moderate-or-severe TR with an AUC of 0.951 (0.938-—0.962) and severe TR with an AUC of 0.980 (0.966-—0.988).

Conclusions and Relevance: In this study, ~~We developed~~ an automated pipeline was developed to identify clinically significant TR with excellent performance.

Commented [EH13]: See prior query re: this terminology.

Commented [EH14]: I'd recommending expanding this section with another sentence. What need does this innovation fill, and what are its clinical implications?

Introduction

Accurate and reliable assessment of tricuspid regurgitation (TR) severity remains an ongoing challenge. Once considered a benign consequence of co-existing heart disease, ~~tricuspid regurgitation~~TR has received more recent recognition as an independent risk factor for morbidity and mortality.[1-3] While recent findings highlight the need for earlier diagnosis and monitoring, TR can often be asymptomatic and present without auscultation findings.[4] With high temporal resolution, transthoracic echocardiography (TTE) is the most common test of choice for characterizing TR;~~,~~ however, accurate diagnosis requires expert assessment, as there is significant intra-observer variability.[5,6] As new therapeutic options for treating TR, like percutaneous repair, emerge,[7,8] early and accurate diagnosis has become more important.

Recent advances in computer vision and artificial intelligence (AI) have enabled precision phenotyping of structure and function in cardiac ultrasound.[9] AI applied to echocardiography can precisely estimate wall thickness~~,~~[10] and assess mitral regurgitation severity~~,~~[11] and left ventricular ejection fraction (LVEF),[12,13] as well as detect cardiac

amyloidosis,[10,14] hypertrophic cardiomyopathy~~HCM~~,[15] and diastolic dysfunction.[16] Application of deep learning for ~~tricuspid regurgitation~~TR has lagged behind, with most machine learning approaches using structured tabular data to characterize and prognosticate TR rather than evaluating the underlying images themselves.[17-20] AI guidance has been developed for both image acquisition and interpretation,[12,21] and given the increasing prevalence of TR in an aging population with co-morbid heart failure, AI could aid in TR screening and surveillance.[22-26]

In the present study, ~~we developed and evaluated~~ a deep learning pipeline was developed and evaluated to detect and assess TR severity from transthoracic echocardiogram studies. Automating the entire process of view selection, identification of ~~tricuspid regurgitation~~TR by color Doppler, and assessment of severity, it was~~we~~ hypothesized that a deep learning approach could assess TR severity with high-throughput automation. This pipeline was evaluated with data from 2~~two~~ geographically distinct sites, including a temporally distinct test cohort ~~distinct~~ separate from the training and validation cohorts (Figure 1). Combined with other echocardiography AI algorithms, like those enabling novices to obtain point of care echocardiographic images,[21] such an approach could be used for serial surveillance and screening of ~~tricuspid regurgitation~~TR.

Commented [EH15]: I made this change just to avoid repeating "distinct," just FYI!

Methods

Study Population and Data Source

Cedars-Sinai Medical Center ~~(CSMC)~~ Cohort

~~Transthoracic echocardiography (~~TTE~~)~~ studies at Cedars-Sinai Medical Center (CSMC) between October ~~0~~4, 2011, and December 31, 2021, were used to train the~~our~~ deep learning pipeline (~~"~~EchoNet-TR~~"~~) for high-throughput TR identification and grading. Studies were initially sourced from Digital Imaging and Communications in Medicine (DICOM) files and underwent de-identification, view classification, and pre-processing into AVI files, as previously described,[25] yielding 2 079 898 videos from 47 312 studies involving~~from~~ 31 708 patients. From these videos, 57 701 apical 4-chamber (A4C) videos with color Doppler across the tricuspid valve were manually curated and used to train a deep learning pipeline for TR phenotyping.

Studies were randomly split on a patient level into training (95%) and internal validation (5%) cohorts to train deep neural networks for TR phenotyping.[27] Identical patient-level splits were maintained training both the TR severity and view classification models. The trained models were evaluated serially as a single pipeline on a held-out temporal test set of 2462 TTE studies (101 455 videos) from 2170 patients receiving care at CSMC between January ~~0~~1, 2022, and June 04, 2022. Patients in the training and validation sets were excluded from the training set.

Stanford Healthcare ~~(SHC)~~ Cohort

The pipeline was evaluated on 5549 studies (containing a total of 278 377 videos) from Stanford Healthcare (SHC)'s high-volume echocardiography lab. The automated view classification pipeline was compared with manual curation of videos within those studies to evaluate specificity. All videos identified by the view classifier were used for downstream TR severity model validation. Model output was compared with TR severity determined by expert cardiologists from the clinical reports. This study was approved by the i~~I~~nstitutional ~~R~~review b~~B~~oards at ~~Cedars-Sinai Medical Center~~CSMC and SHC~~tanford Healthcare~~. Informed consent was waived, as the study involved secondary analysis of existing data without patient contact. Reporting of study results is consistent with guidelines put forth by Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI) reporting guidelines (eTable 12 in Supplement 1).[30,31]

Commented [EH16]: I've moved this info to where it normally appears in JAMA studies, and I've added a Supplement callout where this checklist appears.

AI Model Training

The model pipeline consisted of a view classifier capable of detecting A4C videos with color Doppler across the tricuspid valve from full echocardiographic studies and a TR severity classification model. The PyTorch Lightning deep learning framework was used to train deep learning models. Video-based convolutional neural networks of the R(2 + 1)D architecture were used for view classification and TR severity assessment.[28] The view classification model was initialized with random weights, while the TR severity model was initialized with weights from Echo-Net--Dynamic.[27] Both models were trained using a cross-entropy loss function for up to 100 epochs, an Adam~~DAM~~ optimizer, an initial learning rate of 1^{e-2}, and a batch size of 24 on

Commented [EH17]: Is this nomenclature with the superscript correct? Apologies, I'm unfamiliar.

an NVIDIA GeForce RTX 3090 graphics processing unit~~GPU~~ (NVIDIA Corporation). Early stopping was performed based on the validation loss.

The view classifier was trained using the 57 701 manually curated A4C videos with color Doppler across the tricuspid valve as cases and 421 679 controls from the same studies. Controls consisted of any videos that were not A4C videos with color Doppler across the tricuspid valve and included videos of both A4C and other echocardiographic views (both with and without color Doppler information). The TR severity model was trained using 57 701 manually curated videos, which consisted of 14 318 videos with no TR, 16 507 videos with mild TR, 19 820 videos with moderate TR, and 7056 videos with severe TR. TR severity for each study was determined based on the clinical echocardiographic reports from CSMC's high-volume echocardiography lab, where severity was assessed in accordance with American Society of Echocardiography guidelines.[29] When TR was characterized as an intermediate category (ie, ~~-~~ "trace to mild," ~~or~~ "mild to moderate," or "moderate to severe"), videos were placed in the more severe category. Studies with concomitant tricuspid stenosis, prosthetic valves, and heart failure were also included in both training and validation datasets.

Commented [EH18]: Please confirm these quoted passages are reproduced verbatim (including punctuation) from the study materials. For example, in the Abstract, this was either "mild-to-moderate" or "mild-or-moderate."

Statistical Analysis

The pipeline was then evaluated on 2~~two~~ test sets not seen during model training: 2462 studies obtained at CSMC in 2022 (temporally distinct from the training and validation studies and with no patient overlap) and 5549 studies obtained at SHC in 2018. This process is summarized in Figure 1. Confusion matrices and area under the receiver operating characteristic curve (AUC) were used to assess model performance, and statistics related to TR model

Commented [EH19]: See Abstract query re: AUC vs AUROC.

performance were calculated on a study level. When a study had more than 1 ~~one~~ A4C video with Doppler information across the tricuspid valve, predictions across videos were ensembled, and the video that resulted in the greatest predicted TR severity was used for analysis. When multiple videos resulted in the same maximal predicted severity, the video with the highest prediction probability for the given level of TR severity was used. In both the internal and external test sets, AUC, F1 -score, recall (sensitivity), positive predictive value (PPV), and negative predictive value (NPV) were calculated for clinically significant TR, which was defined as greater than moderate TR and severe TR. AUC was also calculated for relevant subsets in the CSMC test cohort. Statistical analysis was performed in Python ~~(~~version 3.8.0 (Python Software Foundation). Confidence intervals were computed via bootstrapping with 10 000 samples. ~~Reporting of study results is consistent with guidelines put forth by CONSORT-AI.[30,31]~~

Subgroup analysis was conducted to assess model performance in patients with different ranges of right and ~~left ventricular ejection fraction~~LVEF, pulmonary artery pressure, associated co-morbidities (≥mild right atrial dilation, ≥mild mitral regurgitation, ≥moderate aortic stenosis, and ≥moderate aortic regurgitation), study characteristics, ~~associated co-morbidities~~ and other clinical characteristics. Echocardiogram study quality was determined by clinicians and extracted from the clinical report. Studies where clinicians commented on technical difficulty, poor study quality, or where 1 ~~one~~ or more major cardiac structure~~s~~ (left ventricle, right ventricle, pulmonary artery, etc~~.~~) were not well visualized were classified as technically difficult.

Error Mode Analysis

TTEs that were overclassified or under-classified by 1 ~~one~~ class were evaluated, as these made up ~~the majority of~~most model errors. Under-classified TTEs were evaluated to gauge if free-text TTE reports mentioned intermediate categories of TR (ie, "mild to moderate" or ~~,~~ "moderate to severe"). Meanwhile, overclassified TTEs where right ventricular systolic pressure (RVSP) and right atrial or right ventricular (RA/RV) pressure gradient were available were compared to those that were correctly classified for differences in disease severity to better understand erroneous classification of the TR severity model. Continuous variables were compared using a Mann-Whitney *U* -test to assess for differences in means.

Model Explainability

Features identified by the TR severity model were evaluated using saliency mapping, generated using the Integrated Gradients method.[32] This method generated a heatmap for every frame of the video, summarized as a final 2-dimensional heatmap generated by using the maximum value along the temporal axis for each pixel location in the video. Pixels brighter in intensity and closer to yellow were more salient to model predictions, while those darker in color were less important to the model's final prediction. When assessing videos with no TR, heatmaps were obtained by taking the maximum of saliency maps for the moderate and severe class output neurons for each pixel location.

Comparison With Cardiac Magnetic Resonance Imaging

Concordance between model predictions, TTE labels, and magnetic resonance imaging (MRI) labels was evaluated. To accomplish this, a held-~~-~~out set of 572 CSMC patients who did not have studies in the training or validation sets was identified. These patients received a total of at least 1~~one~~ echocardiogram study within 180 days of cardiac MRI assessment of TR. When a patient had more than 1~~one~~ TTE within 180 days of cardac MRI, the TTE that was obtained closest in time to the MRI was used for analysis. MRI labels were abstracted from clinical reports, which used quantitative evaluation to stratify TR at high levels and qualitative evaluation at lower levels.

Results

Study Population

Study and patient characteristics for TTEs used to train the view classification and TR severity models are shown in Table 1 and eTable 1 in Supplement 1, where statistics were calculated on a study level. Of XX total patients, XX patients (XX%) were female. In the CSMC training, validation, and test sets, patients had similar characteristics. Meanwhile, the CSMC and SHC test cohorts differed. The SHC test cohort contained lower numbers of studies from patients with moderate TR (5.0% vs 25.0%) and severe TR (4.4% vs 10.0%). Compared with the CSMC cohort, t~~T~~he SHC cohort also had a lower proportion of videos from Black patients (4.5% vs 14.5%)~~,~~ and a higher proportion of videos from Asian patients (25.2% vs 9.8%).

Commented [EH20]: This will mirror what appears in the Abstract; see Abstract query.

View Classifier Performance Across 2~~Two~~ Institutions

On a test set of 2462 TTEs (101 415 videos) from CSMC not seen during model training, the view classifier had an AUC of 1.000 (0.999-1.000) and at a threshold of 0.800 and~~,~~ identified an A4C video with color doppler across the tricuspid valve in 2410 studies in the test set with a sensitivity of 0.979 (0.973-0.985) and a specificity of 1.000 (1.000-~~—~~1.000). To evaluate generalization of the view classification model at a geographically distinct site, ~~we evaluated~~ its performance was evaluated on 5549 studies (278 377 videos) from SHC. In this external validation set, the view classifier picked up at least 1~~one~~ A4C tricuspid Doppler video in 5268 studies for a sensitivity of 0.949 (0.944-~~—~~0.955) and a specificity of 1.000 (0.999-~~—~~1.000).

Commented [EH21]: See Abstract query.

T~~ricuspid R~~egurgitation Severity Model Performance Across 2~~Two~~ Institutions

The TR severity model showed strong performance in TR detection (Figure 2). In the temporally distinct CSMC test set, the model detected at least moderate TR with an AUC of 0.928 (0.913-~~—~~0.943) and detected severe TR with an AUC of 0.956 (0.940-~~—~~0.969). Severe TR was ruled out with an NPV of 0.966 (0.955-~~—~~0.977) and at least moderate TR was excluded with an NPV of 0.893 (0.871-~~—~~0.914). Further information on TR model performance is presented in Table 2~~3~~ and eTable 2 in Supplement 1. Strong model performance was preserved across institutions. In the 2018 SHC cohort, the model identified severe TR with an AUC of 0.980 (0.966-~~—~~0.989) and moderate or ~~/~~severe TR (defined as at least moderate TR) with an AUC of 0.951 (0.938-~~—~~0.962). In this cohort, the model demonstrated an NPV of 0.987 (0.982-~~—~~0.991) for severe TR and an NPV of 0.994 (0.990-~~—~~0.997) for ~~>~~at least moderate TR.

Commented [EH22]: I'm assuming this should be Table 2 rather than 3, as Table 2 is currently uncited in the Table. If this is incorrect, please let me know.

Weighted Cohen~~'s Kappa~~κ, accuracy, and accuracy for different classes of clinically significant TR are shown in eTable 2 in Supplement 1. Similar results were yielded when the latest TTE available for each patient in the test cohort was used, as~~is~~ reported in eTables 3 ~~and eTable~~ 4 in Supplement 1.

Subset Analysis

The TR severity model showed strong performance across test set subgroups (Table 3). In studies with normal, mildly depressed, or moderately or ~~/~~severely depressed right ventricular~~RV~~ function, moderate or ~~/~~severe TR was detected with AUCs of 0.923 (0.902-0.942), 0.904 (0.847-~~—~~0.951), and 0.861 (0.848-~~—~~0.952), respectively. Meanwhile, severe TR was identified with AUCs of 0.962 (0.940-~~—~~0.979), 0.924 (0.872-~~—~~0.966), and 0.882 (0.873-~~—~~0.966), respectively. Performance was similar across different LVEF ranges. The model also performed similarly well in studies from patients with a history of pre~~-~~existing atrial fibrillation, ~~right atrial~~RA dilation, and co~~-~~existing left~~-~~sided valvular heart disease, with moderate or ~~/~~severe TR detection ranging from 0.886 (0.804-~~—~~0.954) to 0.918 (0.848-~~—~~0.973) and severe TR detection ranging from 0.917 (0.975-~~—~~0.952) to 0.961 (0.908-~~—~~0.997).

MRI Comparison

Model predictions of TR severity were compared with assessments using cardiac magnetic resonance (CMR) imaging, another imaging modality capable of assessing valvular function. The MRI comparison cohort was composed of 572 studies from 572 patients. The cohort consisted of 359 studies (62.7%) with no TR, 187 studies (32.7%) with mild TR, 17

studies (2.9%) with moderate TR, and 9 studies (1.1%) with severe TR (eTable 5 in Supplement 1). Model-predicted TR severity showed strong concordance with CMR assessment of TR severity for moderate or severe TR (AUC: 0.896 [0.822-0.948]) and severe TR (AUC: 0.949 [0.845-0.999]; eTable 6 in Supplement 1). Similarly, cardiologist-determined TR severity from echocardiogram studies also agreed with cardiologist-determined severity using MRI for moderate or severe TR (0.820 [0.686-0.966]) and severe TR (0.841 [0.480-0.997]; eTable 7 in Supplement 1). Meanwhile, the difference in AUC between AI model predictions and cardiologist-based echocardiogram prediction vs MRI were significantly different for at least moderate TR (DeLong test, *P* = .02) but not severe TR (DeLong test, *P* = .13; eTable 8 in Supplement 1). Full results are shown in eFigure 1 and eTables 5-8 in Supplement 1.

Commented [EH23]: *P* values alone are not sufficient; please provide the numbers that were compared to produce the values.

Error Mode Analysis

Moderate and severe TTEs underclassified by 1~~one~~ class were most often ~~"~~mild to moderate~~"~~ (84.17%) or moderate to severe (68.66%; eTable 9 in Supplement 1). Meanwhile, only 2.2% of underclassified TTEs were noted as "Trace/trivial to Mild." Overclassified TTEs (eTable 10 in Supplement 1), which had associated RVSP and RA/RV pressure gradients, showed significantly increased and mean (SD) RA/RV pressure gradients for no TR (21.75 [~~±~~ 8.72] vs 19.44 [~~±~~13.44]; *P* < .01) and mild TR (30.35 [~~±~~9.10] vs 26.96 [~~±~~14.36];; *P* < .001~~1e-4~~) ~~TR~~ compared to their correctly classified counterparts. Similar results were seen for mean (SD) RVSP for no TR (27.08 [~~±~~8.55] vs 21.80 [~~±~~6.02];; *P* < .001~~1e-5~~), mild TR (34.98 [~~±~~10.23] vs 30.93 [~~±~~9.88];; *P* = .02), and moderate TR (50.43 [~~±~~17.59] vs 46.12 [~~±~~28.43];;

Commented [EH24]: Please confirm this is reproduced verbatim from the study materials, including capitalization, which seems a bit odd here.

Commented [EH25]: There's something unclear in the grammar of this sentence, and I'm not quite sure what this section is saying. Would you mind clarifying?

Commented [EH26]: JAMA presents P values smaller than .001 as <.001.

P = .03) cases that were overclassified. Distributions and means of each set are shown in eFigure 2 in Supplement 1.

Model Explainability

The Integrated Gradients method was used to create saliency maps that identified regions of interest in each video contributing the most to detection of TR severity (eFigure 3 in Supplement 1).[32] Saliency maps for the TR severity model demonstrated that the clinically relevant imaging features of TR were important for model predictions, with activation signal localizing to pixels in the color Doppler window and primarily highlighting the TR jet, indicating that the model used appropriate~~,~~ physiologic features of TR to make predictions. eVideos 1-4 in Supplement 1 contain frame-by-frame visualizations of saliency maps.

Discussion

The current work presents a comprehensive, automated pipeline capable of characterizing TR from full echocardiogram studies. In a temporally distinct test cohort, the AI automation pipeline demonstrated strong performance in identifying cases of severe and ~~>~~at least moderate TR, with an AUC greater than ~~>~~0.928 and an NPV greater than ~~>~~0.89. The EchoNet-TR algorithm generalized to thousands of studies from a large, geographically distinct cohort with similar performance. While further work remains, these~~our~~ results present a deep learning model that could aid in screening for clinically significant TR. With open-source code and weights, this project can serve as the foundation for future prospective evaluation of AI-assisted workflows in echocardiography.

Commented [EH27]: This could be an excellent sentence to summarize for the Conclusions and Relevance section.

In AI applied to echocardiography, the right heart is still under-represented. While prior work has shown AI's ability to characterize LVEF, left ventricular hypertrophy LVH, and left-sided valvular lesions,[10,19,32,33,19,32] there has been little work on TR. The present work applies a state-of-the-art video-based architecture for TR detection on a large training dataset and shows that such an approach trains a generalizable AI model for the assessment of regurgitant severity. Despite relying on a single salient view, EchoNet-TR reliably matches expert cardiologists' assessments. Despite such limited information, the performance is comparable to clinical variation across clinicians and across modalities. The current pipeline focusing on the A4C view alone suggests there is richness of ancillary information (right atrial RA enlargement, right ventricular RV systolic function, etc.) that augments the assessment of TR severity, even without 3-D information. This potential use of information mimics clinical decision--making, wherein cardiologists often integrate conflicting metrics from different views and ancillary information to arrive at a final TR severity.

Limitations

There are a few limitations worth considering. While PPV for at least ≥moderate TR was lower in the SHC cohort, the lower population prevalence of such ≥Moderate TR cases (10% vs 35%) was the likely driver, as specificity (0.870 [(0.860-0.880])) was preserved. Similarly, despite weighted Cohen's κ kappa and exact accuracy being lower than other metrics of model performance, ≥at least moderate TR and severe TR were still detected with accuracy of more than ≥0.824 (0.802—0.846) in both test cohorts, indicating that higher misclassification rates of mild TR and no TR were responsible for the lower exact accuracy.

Additionally, EchoNet-TR was trained on study-level clinician assessments of TR severity. TR evaluation is subject to inter-observer variability, and although the large scale of data used here could attenuate some of this variability, training and evaluating models on these labels haves the inherent risk of propagating this variability, especially as the large scale of data in this study did not allow for manual adjudication prior to model training. Future work can focus on more nuanced evaluation of disease, like automatic quantification of regurgitant volume, valve leaflet thickness, orifice area, etc.

A few notable findings with regard to RV function stand out in reviewing cases misclassified by this study's model. Studies that were overclassified by 1 level of severity showed a statistically significant difference in RA/RV pressure gradient and RVSP than those correctly classified, possibly suggesting more severe disease and highlighting possible subtle nuances in disease severity detected by the algorithm. This is also important when considering the algorithm's limitation in differentiating between low levels of TR (no TR and mild TR). Similarly, in the internal test set, most under-classification errors in moderate and severe studies occurred in studies that belonged to intermediate levels of severity, reflecting possible limitation in the use of 4 severity categories or inter-observer variability during classification (eTable 9 in Supplement 1).

Finally, while echocardiography remains first-line, multimodal TR characterization is of interest. The algorithm's decreased performance on CMR-based labels (eFigure 4 and eTables 5-8 in Supplement 1) highlights the imperfect multimodal concordance, even by expert readers.[34,35] However, the authors caution against the notion that CMR should be the

criterion~~gold~~ standard in ~~in~~ assessing severity of TR. Further research on TR characterization across modalities is required~~,~~ that~~however~~ recognizes~~ing~~ that CMR has lower temporal resolution,[6] greater slice thickness,[6] ~~and~~ requires labor intensive segmentation, and relies on software that is newer, proprietary, and less validated.[36]

Conclusions

In summary, we introduce EchoNet-TR, a model for TR screening from single-view TTE videos with excellent performance and generalizability. In this study~~doing so~~, we provide a workflow for isolating tricuspid valve color Doppler videos and assessing TR severity with strong AUC ~~and~~ in 2~~two~~ distinct test cohorts. While more work remains, the current algorithm could potentially be used to increase access to TR screening or enable retrospective institutional database review.[21,43]

Article Information

Accepted for Publication: February 7, 2025.

Corresponding Author: David Ouyang, MD, ~~Assistant Professor,~~ Smidt Heart Institute, Cedars-Sinai Medical Center, 127 S San Vicente Blvd, AHSP Pavilion, Ste A3100, Los Angeles, CA 90048 (d~~D~~avid.o~~O~~uyang@cshs.org).

Author Contributions:- Author X had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Vrudhula, Kwan, Siegel, Ouyang.

Commented [EH28]: Please confirm this information is correct and complete. (JAMA omits titles, so I've deleted this one.)

Commented [EH29]: Please indicate the author or authors (maximum: 2) that this sentence applies to.

Acquisition, analysis, or interpretation of data: Vrudhula, Vukadinovic, Haeffele, Kwan, Berman, Liang, Cheng, Ouyang.

Drafting of the manuscript: Vrudhula, Berman, Ouyang.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Vrudhula, Vukadinovic, Ouyang.

Obtained funding: Cheng, Ouyang.

Administrative, technical, or material support: Haeffele, Kwan, Berman, Cheng, Ouyang.

Supervision: Kwan, Berman, Siegel, Cheng, Ouyang.

~~**Conflict of Interest Disclosures:** D.O. reports support from AstraZeneca Alexion, as well as consulting from EchoIQ, Ultromics, Pfizer, and InVision. A.C.K. reports support consulting fees from InVision.~~

Conflict of Interest Disclosures: Dr Haeffele reported personal fees from Edwards Lifesciences and Laza Medical ~~and personal fees from Edwards Lifescience~~ outside the submitted work. Dr Kwan reported consulting~~personal~~ fees from InV~~v~~ision Medical Technology ~~Corporation Consulting~~ outside the submitted work. Dr Berman reported grants from the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation during the conduct of the study. Dr Cheng reported ~~personal~~ consulting fees from UCB ~~consulting~~ outside the submitted work. Dr Ouyang reported holding equity in~~other from~~ InVision Medical Technology~~Equity~~;~~,~~ grants from Alexion and~~, grants from~~ AstraZeneca;~~,~~ and personal fees from EchoIQ, Pfizer, and ~~personal fees from~~ Ultromics~~,~~ ~~and personal fees from Pfizer~~ outside the submitted work. No other disclosures were reported.

Commented [EH30]: Please confirm this information is correct and complete. Note that this information was taken from the Authorship Forms; if there are important differences between this statement and what was originally included in the manuscript (deleted above), then please let me know what to add.

Funding/Support: At the time of part of this work, Dr Vrudhula was a research fellow supported by the Sarnoff Cardiovascular Research Award. Dr Ouyang reporteds research grants R00 HL157421 and R01HL173526.

Commented [EH31]: What organization or body provided these grants?

~~**Data Sharing Statement:** The dataset of videos and reports used to train and test EchoNet-TR is not publicly available due to its potentially identifiable nature. Our code and model weights are available at https://github.com/echonet/TR.~~

Data Sharing Statement: See Supplement 2~~X~~.

References

1. Chorin E, Rozenbaum Z, Topilsky Y, et al. Tricuspid regurgitation and long-term clinical outcomes. *Eur Heart J Cardiovasc Imaging*. 2020;21(2):157-165. Medline:31544933 doi:10.1093/ehjci/jez216

2. Nishiura N, Kitai T, Okada T, et al. Long-term clinical outcomes in patients with severe tricuspid regurgitation. *J Am Heart Assoc*. 2023;12(1):e025751. Medline:36565178 doi:10.1161/JAHA.122.025751

3. Rao VN, Giczewska A, Chiswell K, et al. Long-term outcomes of phenoclusters in severe tricuspid regurgitation. *Eur Heart J*. 2023;44(21):1910-1923. Medline:36924209 doi:10.1093/eurheartj/ehad133

4. Arsalan M, Walther T, Smith RL II, Grayburn PA. Tricuspid regurgitation diagnosis and treatment. *Eur Heart J*. 2017;38(9):634-638. Medline:26358570 doi:10.1093/eurheartj/ehv487

5. Alexander B, Datta P, Dasari S, et al. Interobserver variation in echocardiographic measurements and grading of tricuspid regurgitation based on a novel web-based assessment environment. *J Cardiothorac Vasc Anesth*. 2022;36(9):3501-3508. Medline:35595583 doi:10.1053/j.jvca.2022.03.025

6. Hahn RT, Thomas JD, Khalique OK, Cavalcante JL, Praz F, Zoghbi WA. Imaging assessment of tricuspid regurgitation severity. *JACC Cardiovasc Imaging*. 2019;12(3):469-490. Medline:30846122 doi:10.1016/j.jcmg.2018.07.033

7. Arnold SV, Hahn RT, Thourani VH, et al; TRISCEND II Pivotal Trial Investigators. Quality of life after transcatheter tricuspid valve replacement: 1-year results from TRISCEND II pivotal trial. *J Am Coll Cardiol*. 2025;85(3):206-216. Medline:39480380 doi:10.1016/j.jacc.2024.10.067

8. Hahn RT, Makkar R, Thourani VH, et al; TRISCEND II Trial Investigators. Transcatheter valve replacement in severe tricuspid regurgitation. *N*

Engl J Med. 2025;392(2):115-126. Medline:39475399 doi:10.1056/NEJMoa2401918

9. Wehbe RM, Katsaggelos AK, Hammond KJ, et al. Deep learning for cardiovascular imaging: a review. *JAMA Cardiol*. 2023;8(11):1089-1098. Medline:37728933 doi:10.1001/jamacardio.2023.3142

10. Duffy G, Cheng PP, Yuan N, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA Cardiol*. 2022;7(4):386-395. Medline:35195663 doi:10.1001/jamacardio.2021.6059

11. Vrudhula A, Duffy G, Vukadinovic M, Liang D, Cheng S, Ouyang D. High-throughput deep learning detection of mitral regurgitation. *Circulation*. 2024;150(12):923-933. Medline:39129623 doi:10.1101/2024.02.08.24302547

12. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*. 2020;580(7802):252-256. Medline:32269341 doi:10.1038/s41586-020-2145-8

13. He B, Kwan AC, Cho JH, et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. *Nature*. 2023;616(7957):520-524. Medline:37020027 doi:10.1038/s41586-023-05947-3

14. Goto S, Mahara K, Beussink-Nelson L, et al. Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. *Nat Commun*. 2021;12(1):2726. Medline:33976142 doi:10.1038/s41467-021-22877-8

15. Sangha V, Dhingra LS, Oikonomou E, et al. Identification of hypertrophic cardiomyopathy on electrocardiographic images with deep learning. *medRxiv*. Preprint published online December 28, 2023. doi:10.1101/2023.12.23.23300490

16. Lee E, Ito S, Miranda WR, et al. Artificial intelligence-enabled ECG for left ventricular diastolic function and filling pressure. *NPJ Digit Med*. 2024;7(1):4. Medline:38182738 doi:10.1038/s41746-023-00993-7

17. Zhao C, Chan NF, Chan RNC, Lee APW. Unsupervised machine learning in echocardiography of functional tricuspid regurgitation. *J Am Coll Cardiol*. 2024;83(13)(suppl):2152. doi:10.1016/S0735-1097(24)04142-1

18. Chorin E, Topilsky Y. The synergy between tricuspid regurgitation and machine learning. *Eur Heart J*. 2023;44(21):1924-1926. Medline:36928709 doi:10.1093/eurheartj/ehad140

19. Deb B, Scott C, Pislaru SV, et al. Machine learning facilitates the prediction of long-term mortality in patients with tricuspid regurgitation. *Open Heart*. 2023;10(2):e002417. Medline:38011995 doi:10.1136/openhrt-2023-002417

20. Heitzinger G, Spinka G, Koschatko S, et al. A streamlined, machine learning-derived approach to risk-stratification in heart failure patients with secondary tricuspid regurgitation. *Eur Heart J Cardiovasc Imaging*. 2023;24(5):588-597. Medline:36757905 doi:10.1093/ehjci/jead009

21. Narang A, Bae R, Hong H, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. *JAMA Cardiol*. 2021;6(6):624-632. Medline:33599681 doi:10.1001/jamacardio.2021.0185

22. Nkomo VT, Gardin JM, Skelton TN, Gottdiener JS, Scott CG, Enriquez-Sarano M. Burden of valvular heart diseases: a population-based study. *Lancet*. 2006;368(9540):1005-1011. Medline:16980116 doi:10.1016/S0140-6736(06)69208-8

23. Martin AC, Bories MC, Tence N, et al. Epidemiology, pathophysiology, and management of native atrioventricular valve regurgitation in heart

failure patients. *Front Cardiovasc Med*. 2021;8:713658. Medline:34760937 doi:10.3389/fcvm.2021.713658

24. Avierinos JF, Tribouilloy C, Grigioni F, et al; Mitral regurgitation International DAtabase (MIDA) Investigators. Impact of ageing on presentation and outcome of mitral regurgitation due to flail leaflet: a multicentre international study. *Eur Heart J*. 2013;34(33):2600-2609. Medline:23853072 doi:10.1093/eurheartj/eht250
25. Grave C, Tribouilloy C, Tuppin P, et al. Fourteen-year temporal trends in patients hospitalized for mitral regurgitation: the increasing burden of mitral valve prolapse in men. *J Clin Med*. 2022;11(12):3289. Medline:35743354 doi:10.3390/jcm11123289
26. Mitchell E, Walker R. Global ageing: successes, challenges and opportunities. *Br J Hosp Med (Lond)*. 2020;81(2):1-9. Medline:32097073 doi:10.12968/hmed.2019.0377
27. Zhang J, Gajjala S, Agrawal P, et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation*. 2018;138(16):1623-1635. Medline:30354459 doi:10.1161/CIRCULATIONAHA.118.034338
28. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. *IEEE Xplore*.

Preprint published online December 16, 2018. doi:10.1109/CVPR.2018.00675

29. Zoghbi WA, Adams D, Bonow RO, et al. Recommendations for noninvasive evaluation of native valvular regurgitation: a report from the American Society of Echocardiography developed in collaboration with the Society for Cardiovascular Magnetic Resonance. *J Am Soc Echocardiogr*. 2017;30(4):303-371. Medline:28314623 doi:10.1016/j.echo.2017.01.007

30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2):336-359. doi:10.1007/s11263-019-01228-7

31. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. Medline:25738806 doi:10.1371/journal.pone.0118432

32. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *arXiv*. Preprint published online March 4, 2017. doi:10.48550/arXiv.1703.01365

33. Holste G, Oikonomou EK, Mortazavi BJ, et al. Severe aortic stenosis detection by deep learning applied to echocardiography. *Eur Heart J*. 2023;44(43):4592-4604. Medline:37611002 doi:10.1093/eurheartj/ehad456

34. Sularz A, Negm AS, Chavez-Ponce A, et al. Prospective quantification of tricuspid regurgitation with echocardiography vs 4D flow cardiac magnetic resonance. *medRxiv*. Preprint published online October 13, 2024. doi:10.1101/2024.10.11.24315310

35. Zhan Y, Senapati A, Vejpongsa P, Xu J, Shah DJ, Nagueh SF. Comparison of echocardiographic assessment of tricuspid regurgitation against cardiovascular magnetic resonance. *JACC Cardiovasc Imaging*. 2020;13(7):1461-1471. Medline:32199838 doi:10.1016/j.jcmg.2020.01.008

36. Suinesiaputra A, Sanghvi MM, Aung N, et al. Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results. *Int J Cardiovasc Imaging*. 2018;34(2):281-291. Medline:28836039 doi:10.1007/s10554-017-1225-9

37. Cohn JN, Tognoni G; Valsartan Heart Failure Trial Investigators. A randomized trial of the angiotensin-receptor blocker valsartan in

chronic heart failure. *N Engl J Med*. 2001;345(23):1667-1675. Medline:11759645 doi:10.1056/NEJMoa010713

38. Pitt B, Zannad F, Remme WJ, et al; Randomized Aldactone Evaluation Study Investigators. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. *N Engl J Med*. 1999;341(10):709-717. Medline:10471456 doi:10.1056/NEJM199909023411001

39. Poole-Wilson PA, Swedberg K, Cleland JGF, et al; Carvedilol Or Metoprolol European Trial Investigators. Comparison of carvedilol and metoprolol on clinical outcomes in patients with chronic heart failure in the Carvedilol Or Metoprolol European Trial (COMET): randomised controlled trial. *Lancet*. 2003;362(9377):7-13. Medline:12853193 doi:10.1016/S0140-6736(03)13800-7

40. McMurray JJV, Packer M, Desai AS, et al; PARADIGM-HF Investigators and Committees. Angiotensin-neprilysin inhibition versus enalapril in heart failure. *N Engl J Med*. 2014;371(11):993-1004. Medline:25176015 doi:10.1056/NEJMoa1409077

41. Solomon SD, McMurray JJV, Claggett B, et al; DELIVER Trial Committees and Investigators. Dapagliflozin in heart failure with mildly reduced or

Commented [EH32]: This reference through reference 42 aren't cited in the text; please indicate where these sources should be cited, or they'll be deleted from the manuscript.

Commented [EH33]: This reference through reference 42 aren't cited in the text; please indicate where these sources should be cited, or they'll be deleted from the manuscript.

Commented [EH34]: This reference through reference 42 aren't cited in the text; please indicate where these sources should be cited, or they'll be deleted from the manuscript.

Commented [EH35]: This reference through reference 42 aren't cited in the text; please indicate where these sources should be cited, or they'll be deleted from the manuscript.

preserved ejection fraction. *N Engl J Med*. 2022;387(12):1089-1098. Medline:36027570 doi:10.1056/NEJMoa2206286

42. Sorajja P, Whisenant B, Hamid N, et al; TRILUMINATE Pivotal Investigators. Transcatheter repair for patients with tricuspid regurgitation. *N Engl J Med*. 2023;388(20):1833-1842. Medline:36876753 doi:10.1056/NEJMoa2300525

43. Chiu IM, Lin CR, Yau FFF, et al. Use of a deep-learning algorithm to guide novices in performing focused assessment with sonography in trauma. *JAMA Netw Open*. 2023;6(3):e235102. Medline:36976564 doi:10.1001/jamanetworkopen.2023.5102

Figure 1. Cedars-Sinai Medical Center (CSMC) and Stanford Dataset Isolation

A total of 57 701 apical 4-chamber~~A4C~~ videos were manually curated from 47 312 studies with varying tricuspid regurgitation (TR) severity and used to train deep learning models for view classification and TR severity stratification. A pipeline consisting of these models was then benchmarked on a temporally distinct cohort of 2462 studies from CSMC and a geographically distinct cohort of 5549 studies from Stanford Healthcare~~SHC~~.

Figure 2. Model Performance Across Severity and Institution

A, Receiver operating characteristic ~~(ROC)~~ curves for detection of s~~S~~evere or ≥at least m~~M~~oderate tricuspid regurgitation (TR) at Cedars-Sinai Medical Center (CSMC) and

Commented [EH36]: This reference through reference 42 aren't cited in the text; please indicate where these sources should be cited, or they'll be deleted from the manuscript.

Commented [EH37]: This reference through reference 42 aren't cited in the text; please indicate where these sources should be cited, or they'll be deleted from the manuscript.

Stanford Healthcare (SHC). The ~~">"~~ at least m~~M~~oderate~~"~~ category (≥moderate) included moderate, moderate to severe, and severe TR. ~~B and C,~~ TR c~~C~~lassification is shown on test set videos from CSMC (B) and SHC~~Stanford~~ (C)~~, respectively~~. Confusion matrix colormap values were scaled based on the proportion of actual disease cases in each class that were predicted in each possible disease category. This was done to allow for relative comparison of model performance across disease classes (n~~N~~one, m~~M~~ild, m~~M~~oderate, and s~~S~~evere), given class imbalance. Statistics and confusion matrices are reported on a study level. When a study had multiple TR Doppler videos, the video with the maximum predicted TR severity was used. When multiple videos lead to the same maximum severity model prediction, the video with the highest prediction probability for that severity of TR was used. AUC indicates area under the curve.

Table 1. Model Derivation Cohort and Test Cohort Characteristics

Characteristic	Cohorts, No. (%)			
	Derivation cohort		Test cohort	
	Training	Validation	CSMC	SHC
Patients	30 125 (88.92)	1583 (4.67)	2170 (6.41)	5014
Studies, No.	44 908	2404	2462	5549
View classification model videos, No.	453 787	24 484	101 415	278 377
TR severity model videos, No.	54 787	3080	2914	7233
Sex				
Female	XX (XX)	XX (XX)	XX (XX)	XX (XX)
Male	23 552 (53.0)	1282 (53.3)	1286 (52.2)	1057 (52.7)[a]
Hypertension	27 743 (61.8)	1477 (61.4)	1295 (52.6)	953 (47.6)[a]
Coronary artery disease	18 983 (42.3)	999 (41.6)	903 (36.7)	721 (36.0)[a]
Atrial fibrillation	14 612 (32.5)	856 (35.6)	515 (20.9)	550 (27.4)[a]
Ejection fraction, mean (SD), %	56.9 (15.4)	57.0 (15.1)	57.2 (15.1)	57.4 (11.0)[a]
Left a~~A~~trial v~~V~~olume i~~I~~ndex, mean (SD), mL/m^2	35.5 (15.7)	36.1 (16.0)	31.8 (15.5)	Not assessed
TR severity				
Control	12 297 (27.4)	627 (26.1)	829 (33.7)	3223 (58.1)
Mild	13 286 (29.6)	710 (29.5)	770 (31.3)	1808 (32.6)
Moderate	14 363 (32.0)	785 (32.7)	617 (25.1)	275 (5.0)

Commented [EH38]: It seems like "Studies" or "Patients" might be the better word here: which is more applicable to this table? Also, re: "No. (%)," I've made guesses about the units of measure for these values, but I'm assuming that some of them will need corrections.

If values in the Table represent number of studies, we could consider moving those numbers into the heading, like "Training (n = 44,908)".

Commented [EH39]: We'll need to add a footnote clarifying why the SHC patients are excluded from the %.

Commented [EH40]: JAMA requires that if data for one sex is included, data should be presented for the other as well. Please provide missing data indicated by XX placeholders.

Commented [EH41]: I'm a bit confused as to how males (23,552) can represent around half (53%) of the training cohort, but that cohort is listed as only having 30,125 patients in it. Would you mind clarifying? (I think clarifying whether values represent studies or patients would be helpful, as I mentioned in the top query.)

Severe	4962 (11.0)	282 (11.7)	246 (10.0)	243 (4.4)
Race				
Asian	3459 (7.7)	192 (8.0)	241 (9.8)	504 (25.2)[a]
Black	6004 (13.4)	288 (12.0)	358 (14.5)	90 (4.5)[a]
White	31 117 (69.3)	1698 (70.6)	1559 (63.3)	1097 (54.8)[a]
Other	4328 (9.6)	226 (9.4)	304 (12.3)	311 (15.5)[a]

Abbreviations: CSMC, Cedars-Sinai Medical Center; SHC, Stanford Healthcare; TR, tricuspid regurgitation.

[a]A total of 2003 studies in the SHC test cohort had information on comorbidities, demographics, and ejection fraction~~EF~~.

Table 2. Model Performance

Site or ~~/~~class	XXXX					
	AUROC	**PPV**	**NPV**	**Recall**	**Specificity**	**F1 ~~-~~score**
CSMC						
≥Moderate TR	0.928 (0.913-0.943)	0.836 (0.799-0.871)	0.893 (0.871-0.914)	0.794 (0.753-0.831)	0.917 (0.904-0.931)	0.814 (0.784-0.843)
Severe TR	0.956 (0.940-0.969)	0.710 (0.622-0.791)	0.966 (0.955-0.977)	0.682 (0.596-0.765)	0.970 (0.963-0.977)	0.696 (0.622-0.759)
SHC						
≥Moderate TR	0.951 (0.938-0.962)	0.427 (0.383-0.466)	0.994 (0.990-0.997)	0.945 (0.914-0.971)	0.870 (0.860-0.880)	0.586 (0.544-0.626)
Severe TR	0.980 (0.966-0.989)	0.699 (0.614-0.780)	0.987 (0.982-0.991)	0.702 (0.616-0.784)	0.986 (0.983-0.990)	0.700 (0.630-0.763)

Abbreviations: AUROC, area under the receiver operating characteristic curve; CSMC, Cedars-Sinai Medical Center; NPV, negative predictive value; PPV, positive predictive value; SHC, Stanford Healthcare; TR, tricuspid regurgitation.

Table 3. Tricuspid Regurgitation (TR) Model Subset Analysis

Subset	~~No. of~~ S~~s~~tudies, No.	TR severity model performance, XXX	
		Moderate/severe	**Severe**
Right ventricular function			
Normal function	1845	0.923 (0.902-0.942)	0.962 (0.940-0.979)
Mildly depressed	285	0.904 (0.847-0.951)	0.924 (0.872-0.966)
Moderately or ~~/~~severely depressed	266	0.861 (0.848-0.952)	0.882 (0.873-0.966)
Left ventricular (LV) systolic function			
LVEF ≥50%	1924	0.930 (0.913-0.947)	0.962 (0.944-0.977)
LVEF <50%	436	0.897 (0.851-0.937)	0.920 (0.876-0.957)

Commented [EH42]: I've reordered racial categories per JAMA style. How was race measured and reported for this study? (eg, drawn from EHR, questionnaire, etc.) Please provide all racial categories assessed.

Commented [EH43]: What other racial subgroups were included in the "Other" category?

Commented [EH44]: See prior comment re: placement of Table 2 callout.

Commented [EH45]: Do these values in parentheses represent 95% CIs? We'll need to include a unit where this placeholder text currently is.

Commented [EH46]: I've rearranged this table a bit to lessen the amount of label text ("moderate/severe" or "severe") in the table itself.

LVEF ≥35%	2105	0.932 (0.915-0.947)	0.960 (0.943-0.974)
LVEF <35%	255	0.860 (0.787-0.923)	0.917 (0.857-0.964)
Atrial fibrillation	499	0.908 (0.869-0.941)	0.917 (0.875-0.952)
Right atrial dilation	505	0.892 (0.847-0.931)	0.920 (0.882-0.952)
Pulmonary artery (PA) pressure			
PA pressure >25 mm Hg	1445	0.892 (0.871-0.916)	0.929 (0.905-0.950)
PA pressure ≤25 mm Hg	622	0.920 (0.835-0.982)	0.998 (0.990-1.000)
PA pressure >35 mm Hg	822	0.850 (0.806-0.890)	0.881 (0.841-0.916)
PA pressure ≤35 mm Hg	1245	0.915 (0.880-0.945)	0.990 (0.975-0.999)
Mitral regurgitation	357	0.896 (0.869-0.922)	0.945 (0.920-0.966)
Aortic stenosis	155	0.918 (0.848-0.973)	0.959 (0.901-0.996)
Aortic regurgitation	144	0.886 (0.804-0.954)	0.961 (0.908-0.997)
Study quality	253	0.927 (0.826-0.989)	0.982 (0.932-1.000)

Abbreviation: LVEF, left ventricular ejection fraction.

Supplement 1. eTable 1. TR Severity Model Derivation Cohort and Test Cohort Characteristics

eTable 2. Cohen's Kappa and Accuracy Analysis

eTable 3. Patient Level Model Performance

eTable 4. Patient Level Analysis of Accuracy & Cohen's Kappa

eTable 5. MRI Cohort Characteristics

eTable 6. Echocardiogram Severity vs MRI Severity

eTable 7. Model Predicted Severity vs MRI Severity

eTable 8. DeLong Test Results

eTable 9. Underclassification Error Mode Analysis

eTable 10. Overclassified Cases - RA/RV Gradient

eTable 11. Overclassified Cases - RVSP

eTable 12. CONSORT-AI Checklist

eFigure 1. Model Predictions, MRI Labels, and TTE Labels

eFigure 2. Error Mode Analysis of Overclassified Cases

eFigure 3. Saliency Map Visualization for TR Classification Models

eVideo 1. CSMC Severe TR Saliency Map

eVideo 2. CSMC No TR Saliency Map

eVideo 3. SHC Severe TR Saliency Map

eVideo 4. SHC No TR Saliency Map

Supplement 2. Data Sharing Statement